 

Let's return to a topic we had in MATH 340.

$$X_1, \ldots, X_n \overset{iid}{\sim} \rho N(\theta_0, \sigma_0^2) + (1-\rho) N(\theta_1, \sigma_1^2) = \sum_{m=1}^{M} \rho_m N(\theta_m, \sigma_m^2)$$

when $M=2$

This was called a mixture model. Parameters? $\{\rho, \theta_0, \sigma_0^2, \theta_1, \sigma_1^2\}$

$$f(\vec{X}; \rho, \theta_0, \sigma_0^2, \theta_1, \sigma_1^2) = \mathcal{L}(\rho, \theta_0, \sigma_0^2, \theta_1, \sigma_1^2; \vec{X}) = \prod_{i=1}^{n} \rho \frac{1}{\sqrt{2\pi\sigma_0^2}} \cdots\cdots + (1-\rho) \cdots$$

MLE's?

$$\ell(\rho, \theta_0, \sigma_0^2, \theta_1, \sigma_1^2; \vec{X}) = \sum_{i=1}^{n} \ln(\rho \cdots\cdots$$

Very difficult to take derivatives and set $=0$ to solve. As $M$ increases, even more difficult.

Enter "data augmentation" (1987)

What if we knew

$X_1$ belonged to $N(\theta_1, \sigma_1^2)$, $X_2$ belonged to $N(\theta_2, \sigma_2^2)$, $X_3 \ldots$ ?

let $I_i = \mathbb{1}_{i\text{th observation "comes from" } N(\theta_0, \sigma_0^2)}$

$\Rightarrow I_1, \ldots, I_n \overset{iid}{\sim} \text{Bern}(\rho)$ "data augmentation" "add" "new data", the $I_i$'s.

$$f(\vec{X}, \vec{I} \mid \rho, \theta_0, \sigma_0^2, \theta_1, \sigma_1^2) = f(\vec{X} \mid \rho, \theta_0, \sigma_0^2, \theta_1, \sigma_1^2, \vec{I})\, P(\vec{I} \mid \rho, \theta_0, \sigma_0^2, \theta_1, \sigma_1^2)$$

$$= \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{2\sigma_0^2}(X_i - \theta_0)^2}\right)^{I_i} \left(\frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2\sigma_1^2}(X_i - \theta_1)}\right)^{1-I_i} \prod_{i=1}^{n} \rho^{I_i}(1-\rho)^{1-I_i}$$

HW: verify $f(\vec{x}|\ldots) = \sum_{S_{\vec{I}}} f(\vec{x}, I|\theta)$

How does this help? We can now find MLE's!

$$\ell(\rho, \theta_0, \sigma_0^2, \theta_1, \sigma_1^2; \vec{x}, \vec{I}) = \sum_{i=1}^{n} \ln(\quad\quad\quad)$$

$$= \sum_{i=1}^{n} I_i\left(-\tfrac{1}{2}\ln(2\pi) - \tfrac{1}{2}\ln(\sigma_0^2) - \tfrac{1}{2\sigma_0^2}(x_i-\theta_0)^2\right)$$
$$+(1-I_i)\left(-\tfrac{1}{2}\ln 2\pi - \tfrac{1}{2}\ln(\sigma_1^2) - \tfrac{1}{2\sigma_1^2}(x_i-\theta_1)^2\right)$$
$$+ I_i\ln(\rho) + (1-I_i)\ln(1-\rho)$$

$$= -\tfrac{1}{2}\ln(2\pi)\sum I_i - \tfrac{1}{2}\ln(\sigma_0^2)\sum I_i - \tfrac{1}{2\sigma_0^2}\sum(x_i-\theta_0)^2 I_i$$
$$-\tfrac{1}{2}\ln(2\pi)\sum(1-I_i) - \tfrac{1}{2}\ln(\sigma_1^2)\sum(1-I_i) - \tfrac{1}{2\sigma_1^2}\sum(x_i-\theta_1)^2(1-I_i)$$
$$+\ln(\rho)\sum I_i + \ln(1-\rho)\sum(1-I_i)$$

let $n_0 = \sum I_i$

let $n_1 = \sum 1-I_i$

$$\frac{\partial \ell}{\partial \theta_0} = -\frac{1}{2\sigma_0^2}\sum\frac{\partial}{\partial\theta_0}(x_i-\theta_0)^2 I_i = -\frac{1}{2\sigma_0^2}\frac{\partial}{\partial\theta_0}\left[\sum x_i^2 I_i - 2\theta_0\sum x_i I_i + \theta_0^2\sum I_i\right] \overset{set}{=} 0$$

$$\Rightarrow -2\sum x_i I_i + 2\theta_0\sum I_i = 0 \Rightarrow \hat{\theta}_0^{mle} = \frac{\sum x_i I_i}{n_0}, \quad \hat{\theta}_1^{mle} = \frac{\sum x_i(1-I_i)}{n_1} \quad \underline{\underline{HW}}$$

$$\frac{\partial \ell}{\partial \sigma_0^2} = -\frac{\sum I_i}{2\sigma_0^2} + \frac{\sum(x_i-\theta_0)^2 I_i}{2(\sigma_0^2)^2} \overset{set}{=} 0 \Rightarrow -\sum I_i + \frac{\sum(x_i-\theta_0)^2 I_i}{\sigma_0^2} = 0 \Rightarrow \hat{\sigma}_0^{2\,mle} = \frac{\sum(x_i-\theta_0)^2 I_i}{n_0}$$

$$\hat{\sigma}_1^{2\,mle} = \frac{\sum(x_i-\theta_0)^2(1-I_i)}{n_1}$$

$$\frac{\partial \ell}{\partial \rho} = \frac{n_0}{\rho} - \frac{n_1}{1-\rho} \overset{set}{=} 0 \Rightarrow n_0(1-\rho) = n_1\rho \Rightarrow n_0 - n_0\rho = n_1\rho \Rightarrow n_0 = (n_0+n_1)\rho$$

$$\Rightarrow \hat{\rho}^{mle} = \frac{n_0}{n_0+n_1}$$

Okay great... who cares? I don't know $I_1, \dots, I_n$! So all of this is useless, but a fun math exercise!

What if we can estimate the $I_i$'s?

since $I_i$ is Bernoulli

$$\text{let } \hat{\hat{I}}_i := E[I_i \mid \theta_0, \sigma_0^2, \theta_1, \sigma_1^2, \rho, X_i] = P(\underbrace{I_i = 1}_{A} \mid \underbrace{\theta_0, \sigma_0^2, \theta_1, \sigma_1^2, \rho}_{C}, \underbrace{X_i}_{B})$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(A|B,C) = \frac{P(B|A,C)\,P(A|C)}{P(B|C)}$$

$$= \frac{f(X_i, I_i = 1 \mid \theta_0, \sigma_0^2, \theta_1, \sigma_1^2, \rho)}{f(\ldots)}$$

$$= \frac{f(X_i \mid \theta_0, \sigma_0^2, \theta_1, \sigma_1^2, I_i = 1)\, \overbrace{P(I_i = 1 \mid \theta_0, \sigma_0^2, \theta_1, \sigma_1^2, \rho)}^{\rho}}{f(X_i \mid \theta_0, \sigma_0^2, \theta_1, \sigma_1^2, \rho)}$$

$$= \frac{\rho \frac{1}{\cancel{\sqrt{2\pi\sigma_0^2}}} e^{-\frac{1}{2\sigma_0^2}(X_i - \theta_0)^2}}{\rho \frac{1}{\cancel{\sqrt{2\pi\sigma_0^2}}} e^{-\frac{1}{2\sigma_0^2}(X_i - \theta_0)^2} + (1-\rho)\frac{1}{\cancel{\sqrt{2\pi\sigma_1^2}}} e^{-\frac{1}{2\sigma_1^2}(X_i - \theta_1)^2}}$$

Why not do the following:

Step 0: Initialize $\hat{\theta}_0 = 0,\ \sigma_0^2 = 1,\ \hat{\theta}_1 = 1,\ \sigma_1^2 = 1,\ \rho = 0.5$

Step 2: Compute $\hat{\theta}_0, \hat{\sigma}_0^2, \hat{\theta}_1, \hat{\sigma}_1^2, \hat{\rho}^{MLE}$ based on $\hat{I}_1, \dots, \hat{I}_n$

Step 1: Compute $\hat{I}_1, \dots, \hat{I}_n$ based on $\hat{\theta}_0, \hat{\sigma}_0^2, \hat{\theta}_1, \hat{\sigma}_1^2, \hat{\rho}$

Step 3: Repeat steps 1, 2 until $\|\vec{\theta}_t - \vec{\theta}_{t-1}\| < \varepsilon$

This is called the Expectation - Maximization Algorithm (EM, 1977)

step 1 (Expectation), step 2 (Maximization)

This is proven to converge to $\hat{\vec{\theta}}^{MLE}$. Also, it's possible to derive the Fisher Information estimates. This gives you asymptotically valid inference via the MLE machinery.

The EM algorithm is very general. Not just for mixture models inference.

Note: easy to upgrade to $M > 2$ $\vec{I}_1, \ldots, \vec{I}_n \overset{iid}{\sim}$ Multinom$(1, \vec{\rho})$ $\boxed{\text{MA HW}}$

Bayesian Approach. Use data augmentation as well. Treat $\vec{I}$ as parameter

let $\underbrace{f(\theta_0) \propto 1, f(\sigma_0^2) \propto \frac{1}{\sigma_0^2}}_{\sigma \text{ Jeffreys}}$, $\underbrace{f(\theta_1) \propto 1, f(\sigma_1^2) \propto \frac{1}{\sigma_1^2}}_{\sigma \text{ Jeffreys}}$, $\underbrace{f(\rho) = U(0,1)}_{\text{Laplace}}$

$f(\vec{I}) \propto 1$

$$f(\theta_0, \sigma_0^2, \theta_1, \sigma_1^2, \rho, \vec{I} \mid \vec{X}) \propto (\sigma_0^2)^{-n/2 - 1} e^{-\frac{1}{2\sigma_0^2} \sum (x_i - \theta_0)^2 I_i} \cdot$$

$$(\sigma_1^2)^{-n/2 - 1} e^{-\frac{1}{2\sigma_1^2} \sum (x_i - \theta_1)^2 (1 - I_i)}$$

$$\rho^{n_0} (1-\rho)^{n_1}$$

Try to make a Gibbs Sampler . . .

$$f(\theta_0 \mid \text{---}) \propto e^{-\frac{1}{2\sigma_0^2}\left(\sum x_i^2 I_i - 2\theta_0 \sum x_i I_i + \theta_0^2 \sum I_i\right)}$$

$$\propto e^{\frac{\sum x_i I_i}{\sigma_0^2} \theta_0 - \frac{n_0}{2\sigma_0^2} \theta_0^2} = e^{a\theta_0 - b\theta_0^2}$$

$$\propto N\left(\frac{\sum x_i I_i}{n_0}, \frac{\sigma_0^2}{n_0}\right)$$

$$f(\theta_1 \mid \text{---}) \propto N\left(\frac{\sum x_i (1 - I_i)}{n_1}, \frac{\sigma_1^2}{n_1}\right)$$

$$f(\sigma_0^2 \mid \text{—}) \propto (\sigma^2)^{-\frac{n_0}{2}-1} e^{-\frac{\sum (x_i-\theta_0)^2 I_i/2}{\sigma_0^2}}$$

$$\propto \text{InvGamma}\left(\frac{n_0}{2}, \frac{\sum (x_i-\theta_0)^2 I_i}{2}\right)$$

$$f(\sigma_1^2 \mid \text{—}) \propto \text{InvGamma}\left(\frac{n_1}{2}, \frac{\sum (x_i-\theta_0)^2 (1-I_i)}{2}\right)$$

$$f(\rho \mid \text{—}) \propto \text{Beta}(n_0+1, n_1+1)$$

$$p(I_i \mid \text{—}) \propto e^{-\frac{1}{2\sigma_0^2}(x_i-\theta_0) I_i} e^{-\frac{1}{2\sigma_1^2}(x_i-\theta_1)(1-I_i)} \rho^{I_i} (1-\rho)^{1-I_i}$$

$$= \left(\rho e^{-\frac{x_i-\theta_0}{2\sigma_0^2}}\right)^{I_i} \left((1-\rho) e^{-\frac{(x_i-\theta_1)}{2\sigma_1^2}}\right)^{1-I_i}$$

$$\propto \text{Bern}\left(\frac{\rho e^{-\frac{x_i-\theta_0}{2\sigma_0^2}}}{\rho e^{-\frac{x_i-\theta_0}{2\sigma_0^2}} + (1-\rho) e^{-\frac{x_i-\theta_1}{2\sigma_1^2}}}\right)$$

We made a Gibbs Sampler!

---

Remember, Stan does not allow for discrete params.

So in Stan, we write the lik without data augmentation